

Adjusting Leverage Scores by Row Weighting: A Practical Approach to Coherent Matrix Completion

Shusen Wang
College of Computer Science
and Technology
Zhejiang University
wss@zju.edu.cn

Tong Zhang
Department of Statistics
Rutgers University
tzhang@stat.rutgers.edu

Zhihua Zhang
Department of Computer
Science and Engineering
Shanghai Jiao Tong University
zhizhua@sjtu.edu.cn

ABSTRACT

Low-rank matrix completion is an important problem with extensive real-world applications. When observations are uniformly sampled from the underlying matrix entries, existing methods all require the matrix to be incoherent. This paper provides the first working method for coherent matrix completion under the standard uniform sampling model. Our approach is based on the weighted nuclear norm minimization idea proposed in several recent work, and our key contribution is a practical method to compute the weighting matrices so that the leverage scores become more uniform after weighting. Under suitable conditions, we are able to derive theoretical results, showing the effectiveness of our approach. Experiments on synthetic data show that our approach recovers highly coherent matrices with high precision, whereas the standard unweighted method fails even on noise-free data.

1. INTRODUCTION

Matrix completion is a well established problem with extensive real-world applications such as collaborative filtering [8, 22], computer vision [14, 25, 11], localization in sensor networks [23, 3], etc. Matrix completion is usually formulated as the penalized nuclear norm minimization model [3, 17] or its variants [24], and can be efficiently solved by many algorithms [30, 19, 29].

Let \mathbf{M} be an $n_1 \times n_2$ matrix with rank $k \ll \min(n_1, n_2)$. We observe only a portion of the entries of \mathbf{M} and seek to recover \mathbf{M} based on the incomplete observations. Let $\Omega \subset [n_1] \times [n_2]$ be an index set such that $(i, j) \in \Omega$ if the (i, j) -th entry of \mathbf{M} is observed, and let $\mathcal{P}_\Omega(\mathbf{M})$ be an $n_1 \times n_2$ matrix with $[\mathcal{P}_\Omega(\mathbf{M})]_{ij} = M_{i,j}$ for all $(i, j) \in \Omega$ and $[\mathcal{P}_\Omega(\mathbf{M})]_{ij} = 0$ for all $(i, j) \notin \Omega$. In order to recover \mathbf{M} based on the partial observation, the cardinality of Ω must be greater than some factor.

It was shown in [4] that under the *uniform sampling model*, that is, each entry of \mathbf{M} is observed independently with the same probability, the sample complexity must be greater than $cnk(\mu + \nu) \log n$ for some constant c in order to exactly recover \mathbf{M} . Here $n = n_1 + n_2$, and $\mu \in [1, \frac{n_1}{k}]$ and $\nu \in [1, \frac{n_2}{k}]$ are the row and column matrix coherence of \mathbf{M} , respectively. When the matrix

coherence is as large as $\mu + \nu = \Theta(\frac{n}{k \log n})$, it is simply impossible to directly complete the matrices.

Some recent work attempts to alleviate the matrix coherence requirements. [16] proposed an active learning approach which allows the row space to be coherent. They use adaptive sampling to select columns followed by accessing all entries in the selected columns. However, their active learning setting is not suitable for most real-world problems because it is usually impossible to access all the missing entries of a column. For example, it is impossible to demand all users to rate one specific item or to pay one user to rate every item in the system.

In another recent work [5], the matrix coherence requirement is eliminated by assuming that each entry be observed independently with probability proportional to the sum of its row and column leverage scores. Obviously, the assumption is much more restrictive than the uniform sampling assumption used in the previous work of [4, 15, 2, 18].

The results of [5] can be used in the reverse direction: one may adjust the leverage scores to align with a given set of observations using appropriate weighting described below. Suppose the (i, j) -th entry of \mathbf{M} is observed with probability p_{ij} . Let $\mathbf{R} \in \mathbb{R}^{n_1 \times n_1}$ and $\mathbf{C} \in \mathbb{R}^{n_2 \times n_2}$ be diagonal weight matrices with positive diagonal entries. Instead of directly completing the matrix $\mathcal{P}_\Omega(\mathbf{M})$, one may first compute \mathbf{R} and \mathbf{C} and then complete the matrix $\mathbf{R}\mathcal{P}_\Omega(\mathbf{M})\mathbf{C} = \mathcal{P}_\Omega(\mathbf{R}\mathbf{M}\mathbf{C})$. The column and row weighting obviously changes the leverage scores of \mathbf{M} . After weighting, if the sum of the i -th row and the j -th column leverage scores of $\mathbf{R}\mathbf{M}\mathbf{C}$ is proportional to p_{ij} , then the dependence of the sample complexity on matrix coherence is eliminated. As a special case, if the observations are uniformly sampled, one needs to find \mathbf{R} and \mathbf{C} such that the leverage scores of $\mathbf{R}\mathbf{M}\mathbf{C}$ are as uniform as possible. This is the idea that motivates our work. We note that none of the previous studies provides any satisfactory approach to find the weight matrices \mathbf{R} and \mathbf{C} ; therefore the focus and the main contribution of this paper is a practical algorithm for computing these matrices.

This work offers the following contributions.

- This paper provides the first practical approach to coherent matrix completion. Our method only makes the uniform sampling assumption, which is standard in the literature; our method can be potentially applied to the non-uniform sampling settings. Before this work, there is no way to completing coherent matrix without adding unrealistic assumptions, e.g. fully observing a number of rows/columns [16], accessing any unobserved entry as one wish [5], etc.
- To complete coherent matrix, the matrix leverage scores must be known or at least approximately estimated. We provide in Theorem 4 a way to estimate the leverage scores of \mathbf{M} from

its incompletely observed entries, and the estimated leverage scores are near their true values with additive-error bound. This result may be of independent interest.

- We derived an ADMM algorithm for solving the weighted nuclear norm minimization model (5). Our ADMM algorithm is efficient on single machine.
- We apply our method to improve another low-rank matrix recovery method called the robust principal component analysis (RPCA) [2]. Experiments on coherent low-rank matrix show that our proposed weighted RPCA exactly recovers the low-rank matrix from heavily noisy observation, whereas the standard RPCA fails on either noisy or noise-free data.

The rest of this paper is organized as follows. Section 2 defines the notation used in this paper. Section 3 formally describe the matrix completion problem and provides an efficient algorithm for solving the weighted nuclear norm minimization problem. Section 4 formulates an optimization model, by solving which the weight matrices can be found and coherent matrices can be completed. Section 5 provides an additive-error perturbation bound for estimating the leverage scores from incompletely observed matrix and applies this technique to perform row weighting. Sections 6 and 7 devise practical algorithms for solving the model formulated in Section 4. Section 8 empirically evaluates our proposed methods on several synthetic datasets. Section 9 applies the proposed row weighting method to improve the robust principal component analysis (RPCA) method. The appendix is available at arXiv:1412.7938, and the MATLAB code is on the first author's home page.

2. NOTATION AND PRELIMINARIES

Let $[m]$ denote the set $\{1, 2, \dots, m\}$. Given a matrix \mathbf{A} , let $\mathbf{a}^{(i)}$ be its i -th row, \mathbf{a}_j be its j -th column, and A_{ij} be its (i, j) -th entry. Let \mathbf{I}_n be the $n \times n$ identity matrix, $\mathbf{0}$ be the all zero vector (or matrix) of the appropriate size, and \mathbf{e}_i be the i -th standard basis whose i -th entry is one and the remaining entries are zero.

Suppose we are given an $n_1 \times n_2$ matrix \mathbf{A} . The singular value decomposition of \mathbf{A} is

$$\begin{aligned} \mathbf{A} &= \sum_i \sigma_i(\mathbf{A}) \mathbf{u}_i \mathbf{v}_i^T = \mathbf{U}_\mathbf{A} \Sigma_\mathbf{A} \mathbf{V}_\mathbf{A} \\ &= \underbrace{\mathbf{U}_{\mathbf{A},k} \Sigma_{\mathbf{A},k} \mathbf{V}_{\mathbf{A},k}}_{:=\mathbf{A}_k} + \mathbf{U}_{\mathbf{A},-k} \Sigma_{\mathbf{A},-k} \mathbf{V}_{\mathbf{A},-k}. \end{aligned}$$

The matrix norms are defined as follows. Let $\|\mathbf{A}\|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$ be the matrix Frobenius norm, $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$ be the spectral norm, $\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A})$ be the matrix nuclear norm, $\|\mathbf{A}\|_1 = \sum_{i,j} |A_{ij}|$ be the matrix ℓ_1 norm, and $\|\mathbf{A}\|_\infty = \max_{i,j} |A_{ij}|$ be the infinity norm. We let the ℓ_0 pseudo-norm $\|\mathbf{A}\|_0$ be the number of nonzero entries of \mathbf{A} .

The row leverage scores of \mathbf{A} according to the best rank k approximation are defined by

$$\mu_i(\mathbf{A}_k) = (\mathbf{U}_{\mathbf{A},k} \mathbf{U}_{\mathbf{A},k}^T)_{ii}.$$

The row cross leverage scores are defined by

$$\mu_{il}(\mathbf{A}_k) = (\mathbf{U}_{\mathbf{A},k} \mathbf{U}_{\mathbf{A},k}^T)_{il},$$

The column leverage scores ν_j and cross leverage scores ν_{jl} are defined similarly by replacing $\mathbf{U}_{\mathbf{A},k}$ by $\mathbf{V}_{\mathbf{A},k}$. The leverage scores and the cross leverage scores satisfy

$$\mu_i(\mathbf{A}_k) \leq 1 \quad \text{and} \quad \sum_{i=1}^{n_1} \mu_i(\mathbf{A}_k) = k, \quad (1)$$

$$\mu_i(\mathbf{A}_k) = \sum_{j=1}^{n_1} \mu_{ij}^2(\mathbf{A}_k). \quad (2)$$

The row matrix coherence of \mathbf{A}_k is defined by

$$\mu(\mathbf{A}_k) = \frac{n_1}{k} \max_i \mu_i(\mathbf{A}_k) \in \left[1, \frac{n_1}{k}\right],$$

and the column coherence ν is defined similarly. In the matrix completion problem with uniform sampling, the idealized leverage scores are $\mu_1 = \dots = \mu_{n_1} = \frac{k}{n_1}$, where the matrix coherence attains its minimum: $\mu = 1$.

In this paper we are particularly interested in the called rank one row weighting, that is, scaling one row by a factor $\sqrt{1-\gamma} \in (0, 1)$. To represent such a row weighting matrix, we define the notation:

$$\mathcal{W}(n_1, i, \gamma) = \mathbf{I}_{n_1} + (\sqrt{1-\gamma} - 1) \mathbf{e}_i \mathbf{e}_i^T, \quad (3)$$

whose the i -th diagonal entry equals to $\sqrt{1-\gamma}$. Suppose we are given the row leverage scores and cross leverage scores of \mathbf{M} , then the (cross) leverage scores of $\mathcal{W}(n_1, i, \gamma)\mathbf{M}$ can be computed in a closed form by the following lemma.

LEMMA 1 ([7]). *Given any $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ and $\gamma \in (0, 1)$, let $\mathbf{W} = \mathcal{W}(n_1, i, \gamma)$ be defined in (3). Then the i -th row leverage score of \mathbf{WM} is*

$$\mu_i(\mathbf{WM}) = \frac{(1-\gamma)\mu_i(\mathbf{M})}{1-\gamma\mu_i(\mathbf{M})} \leq \mu_i(\mathbf{M}),$$

and the j -th ($j \neq i$) leverage scores are

$$\mu_j(\mathbf{WM}) = \mu_j(\mathbf{M}) + \frac{\gamma\mu_{ij}(\mathbf{M})^2}{1-\gamma\mu_i(\mathbf{M})} \geq \mu_j(\mathbf{M}).$$

The cross leverage scores ($j, l \neq i$) are

$$\begin{aligned} \mu_{ij}(\mathbf{WM}) &= \left(1 + \frac{\gamma\mu_i(\mathbf{M})}{1-\gamma\mu_i(\mathbf{M})}\right) \mu_{ij}(\mathbf{M}), \\ \mu_{jl}(\mathbf{WM}) &= (1-\gamma)\mu_{jl}(\mathbf{M}) + \gamma(1-\gamma) \frac{\mu_{ij}(\mathbf{M})\mu_{il}(\mathbf{M})}{1-\gamma\mu_i(\mathbf{M})}. \end{aligned}$$

PROOF. The first two equations are given in Lemma 5 of [7]. The last two equations can be proved directly by the techniques of [7]. \square

Since this kind of row weighting does not change matrix rank, we have

$$\sum_{l=1}^{n_1} \mu_l(\mathbf{WM}) = \text{rank}(\mathbf{WM}) = \text{rank}(\mathbf{M}) = \sum_{l=1}^{n_1} \mu_l(\mathbf{M}).$$

That is, the sum of leverage scores remains constant during row weighting. Therefore, if the i -th row is scale, the decrease in the i -th leverage score turns to the increase in the j -th leverage score for all $j \neq i$.

3. MATRIX COMPLETION

In Section 3.1 we describe the standard nuclear norm minimization model for the matrix completion problem and discuss the sample complexities under different sampling models. In Section 3.2 we introduce the weighted nuclear norm minimization model which is explored in this paper. In Section 3.3 we provide an efficient algorithm for solving the weighted nuclear norm minimization problem; the algorithm is described in Algorithm 1.

3.1 Nuclear Norm Minimization

Let \mathbf{M} be an $n_1 \times n_2$ matrix with rank $k \ll \min(n_1, n_2)$, and let $n = n_1 + n_2$. Given a partial observation $\mathcal{P}_\Omega(\mathbf{M})$, one can solve the following nuclear norm minimization problem to obtain a low-rank matrix \mathbf{L}^* which approximates \mathbf{M} .

$$\min_{\mathbf{L}} \|\mathbf{L}\|_*; \quad \text{s.t. } L_{ij} = M_{ij} \quad \text{for all } (i, j) \in \Omega. \quad (4)$$

This model has been well studied in the literature.

Under the uniform sampling model, [4] showed that $|\Omega| = \mathcal{O}(nk(\mu + \nu)^2 \log^6 n)$ or $|\Omega| = \mathcal{O}(n(\mu + \nu)^4 \log^2 n)$ will be sufficient for the exact recovery of \mathbf{M} with high probability (when entries of \mathbf{M} are not corrupted with noise). Later on, [18] improved the sample complexity to $|\Omega| = \mathcal{O}(\max\{\tau, \mu + \nu\}nk \log^2 n)$, where $\tau = \frac{n_1 n_2}{k} \|\mathbf{U}_{\mathbf{M}, k} \mathbf{V}_{\mathbf{M}, k}^T\|_\infty^2$.

If the observations are non-uniformly sampled, with the (i, j) -th entry of \mathbf{M} sampled according to a probability $p_{ij} \propto \mu_i + \nu_j$, then $|\Omega| = \mathcal{O}(nk \log^2 n)$ will be sufficient for exact recovery [5]. Under this setting, the sample complexity does not depend on the matrix coherence.

3.2 Weighted Nuclear Norm Minimization

A generalization of the standard nuclear norm minimization was proposed in [21, 9] for matrix completion, known as the weighted nuclear norm minimization model:

$$\min_{\mathbf{L}} \|\mathbf{RLC}\|_*; \quad \text{s.t. } L_{ij} = M_{ij} \quad \text{for all } (i, j) \in \Omega, \quad (5)$$

where \mathbf{R} and \mathbf{C} are diagonal matrices.

This model was proposed in [9] to tackle the matrix completion problem when the elements in Ω are not uniformly sampled from $[n_1] \times [n_2]$; that is, M_{ij} is observed with probability p_{ij} which is not necessary that $p_{ij} = \frac{|\Omega|}{n_1 n_2}$. In [9] the authors set $R_{ii}^2 = \sum_j p_{ij} := p_i^R$ and $C_{jj}^2 = \sum_i p_{ij} := p_j^C$ to compensate the unbalance in the observation.

Unfortunately, this kind of weighting does not help completing coherent matrices. Consider a highly coherent matrix whose entries are observed uniformly at random. In this example, the weight matrices used in [9] should be $\mathbf{R} = \sqrt{n_2 p} \mathbf{I}_{n_1}$ and $\mathbf{C} = \sqrt{n_2 p} \mathbf{I}_{n_1}$. Thus, (5) degenerates to (4).

In fact, naively setting \mathbf{R} and \mathbf{C} according to the sampling probabilities $\{p_i^R\}$ and $\{p_j^C\}$ is not the correct way of weighting. Very recently, [5] showed that \mathbf{R} and \mathbf{C} should be chosen such that the leverage scores of \mathbf{RMC} are aligned with the non-uniform observations; that is, $\mu_i(\mathbf{RMC}) + \nu_j(\mathbf{RMC})$ should be proportional to p_{ij} . However, how to compute such weight matrices remains unresolved. We observe that simply setting $R_{ii} = \max\{\mu_i^{-\alpha}(\mathbf{M}), \beta\}$ (and similarly for C_{jj}) and tuning α and β does not work. We thus need to devise a more sophisticated method to find \mathbf{R} and \mathbf{C} , which is the main focus of the paper.

3.3 Solving the Weighted Nuclear Norm Minimization

In practice, since the observation is perturbed by noise, it is better to use the following regularized alternative of (5):

$$\min_{\mathbf{L}} \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{L}) - \mathcal{P}_\Omega(\mathbf{M})\|_F^2 + \lambda \|\mathbf{RLC}\|_*. \quad (6)$$

Let \mathbf{U} and \mathbf{V} be two matrices of sizes $n_1 \times r$ and $n_2 \times r$, respectively. It is well known that when $r = \min\{n_1, n_2\}$, we have that

$$\|\mathbf{RLC}\|_* = \min_{\mathbf{L}=\mathbf{UV}^T} \frac{1}{2} \left\{ \|\mathbf{RU}\|_F^2 + \|\mathbf{CV}\|_F^2 \right\}.$$

Algorithm 1 ADMM for the Weighted Nuclear Norm Minimization.

- 1: **Input:** the partially observed matrix $\mathcal{P}_\Omega(\mathbf{M})$, the parameter λ , diagonal matrices \mathbf{R} and \mathbf{C} .
- 2: Initialize $\mathbf{L}, \mathbf{X}, \mathbf{Y}, \rho$;
- 3: **repeat**
- 4: For all $(i, j) \in \Omega$ update L_{ij} by

$$L_{ij} \leftarrow (1 + \rho R_{ii}^2 C_{jj}^2)^{-1} (M_{ij} + \rho R_{ii} C_{jj} X_{ij} - R_{ii} C_{jj} Y_{ij});$$
- 5: For all $(i, j) \notin \Omega$ update L_{ij} by

$$L_{ij} \leftarrow R_{ii}^{-1} C_{jj}^{-1} (X_{ij} - \rho^{-1} Y_{ij});$$
- 6: $[\mathbf{U}, \Sigma, \mathbf{V}] \leftarrow \text{SVD}(\rho^{-1} \mathbf{Y} + \mathbf{RLC})$;
- 7: Update \mathbf{X} by

$$\mathbf{X} \leftarrow \mathbf{U} [\Sigma - \rho^{-1} \lambda \mathbf{I}]_+ \mathbf{V}^T;$$
- 8: $\mathbf{Y} \leftarrow \mathbf{Y} + \rho(\mathbf{RLC} - \mathbf{X})$;
- 9: **until** converged
- 10: **Output:** \mathbf{L} .

Following the max-margin matrix factorization model [24, 20], [21] proposed the following optimization problem to replace (6):

$$\min_{\mathbf{U}, \mathbf{V}} \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{UV}^T) - \mathcal{P}_\Omega(\mathbf{M})\|_F^2 + \frac{\lambda}{2} (\|\mathbf{RU}\|_F^2 + \|\mathbf{CV}\|_F^2). \quad (7)$$

This model can be solved in the same way as the standard max-margin matrix factorization model by algorithms such as coordinate descent [28], alternating least square [31, 13], stochastic gradient descent [10], etc.

Model (7) naturally admits parallel computing algorithms which are efficient on distributed systems; however, their convergence is usually slow, which leads to poor performance on single machine. In our off-line experiments performed on a single machine, coordinate descent, alternating least square, and stochastic gradient descent for solving (7) all converge slowly. We thus employ a different algorithm to solve (7) based on the alternating direction method with multiplier (ADMM) [1]. The algorithm is described in Algorithm 1 and its derivation is left to Appendix A. Empirically, on a single machine, the proposed ADMM algorithm is significantly faster than the algorithm of [21].

4. COMPUTING THE WEIGHT MATRICES BY OPTIMIZATION

In this section we formulate an optimization model for finding the weight matrices \mathbf{R} and \mathbf{C} . Since row weighting does not affect the column leverage scores, we will discuss only row weighting in the remainder of this paper. The column weight matrix \mathbf{C} can be similarly computed. We propose to minimize the ℓ_q hinge loss function to find the diagonal weight matrix \mathbf{R} :

$$\min_{0 \prec \mathbf{R} \preceq \mathbf{I}_{n_1}} L_{\mathbf{M}, q}(\mathbf{R}) := \left(\sum_{i=1}^{n_1} \max\{\mu_i(\mathbf{RM}) - \mu_i^*, 0\}^q \right)^{\frac{1}{q}}, \quad (8)$$

where $\{\mu_i^*\}$ are defined in (10) below. In the following we will justify this model, devise a coordinate descent algorithm to solve this model, and discuss the choice of q .

4.1 Model Formulation

It was shown in [5] that the sample complexity can be reduced if the leverage scores are aligned with the sampling probabilities of the entries. We seek to find the row and column weight matrices \mathbf{R}

and \mathbf{C} by solving some optimization models such that the leverage scores of \mathbf{RMC} are aligned with the sampling probabilities. We begin with a general non-uniform sampling model.

Let the (i, j) -th entry of \mathbf{M} be observed with probability p_{ij} ($\geq \min\{n_1, n_2\}^{-10}$), and denote

$$p_i^R = \sum_{j=1}^{n_2} p_{ij} \quad \text{and} \quad p_i^C = \sum_{j=1}^{n_1} p_{ij}.$$

It was shown in [5] that the sample complexity $|\Omega|$ does not depend on the matrix coherence parameters provided that

$$p_{ij} \geq \min \left\{ c_0 \frac{(n_1 \mu_i + n_2 \nu_j) \log^2(n_1 + n_2)}{\min\{n_1, n_2\}}, 1 \right\}, \quad (9)$$

where c_0 is constant. Assume that

$$|\Omega| = \sum_{ij} p_{ij} \geq 2c_0 \max\{n_1, n_2\} k \log^2(n_1 + n_2).$$

We can write inequality (9) as

$$c_1 p_{ij} \geq n_1 \mu_i + n_2 \nu_j,$$

summing up the two sides of which w.r.t. i or/and j yields

$$\begin{aligned} c_1 p_i^R &\geq n_2(n_1 \mu_i + k), & c_1 p_j^C &\geq n_1(n_2 \nu_j + k), \\ c_1 &= 2n_1 n_2 k / \sum_{ij} p_{ij}. \end{aligned}$$

Hence, the desired leverage scores are those satisfying

$$\begin{aligned} \mu_i &\leq \frac{2k p_i^R}{\sum_{i=1}^{n_1} p_i^R} - \frac{k}{n_1} := \mu_i^*, \\ \nu_j &\leq \frac{2k p_j^C}{\sum_{j=1}^{n_2} p_j^C} - \frac{k}{n_2} := \nu_j^*. \end{aligned} \quad (10)$$

Notice that μ_i^* and ν_j^* can be less than zero, indicating that (9) does not hold even if the leverage scores are adjusted by row weighting. For such rows or columns, we can abandon them by setting the corresponding weights R_{ii} or C_{jj} to be zero.

Suppose we are given $\mathcal{P}_\Omega(\mathbf{M})$. We can estimate p_i^R and p_j^C by counting the indices in the set Ω , and then compute $\{\mu_i^*\}$ and $\{\nu_j^*\}$ according to (10). We now hope to find the diagonal matrices \mathbf{R} and \mathbf{C} such that $\mu_i(\mathbf{RMC})$ and $\nu_j(\mathbf{RMC})$ are less than or equal to μ_i^* and ν_j^* , respectively. We expect that

$$\mu_i(\mathbf{RM}) = \mu_i(\mathbf{RMC}) \leq \mu_i^* \quad \text{for all } i \in [n_1].$$

Thus, any leverage score of \mathbf{RM} violating the inequality should be penalized. Based on the above discussion, we propose the optimization model (8) to find the weight matrix \mathbf{R} .

As a special case, if the entries of \mathbf{M} are uniformly observed, that is, $p_{ij} = p$ for all i and j , then the optimal row leverage scores are all $\mu_i^* = \frac{k}{n_1}$. The ℓ_q hinge loss function becomes

$$L_{\mathbf{M},q}(\mathbf{R}) := \left(\sum_{i=1}^{n_1} \max \left\{ \mu_i(\mathbf{RM}) - \frac{k}{n_1}, 0 \right\}^q \right)^{\frac{1}{q}}. \quad (11)$$

4.2 Solving Model (8) by Coordinate Descent

In this subsection we propose a coordinate descent algorithm to solve the model (8). Here we assume that we have access to the exact leverage scores; this is obviously unrealistic in the matrix completion problem, but it can help us have a good understanding of our work. In later sections we will adapt this algorithm to solve the real matrix completion problem by estimating the leverage scores from the incomplete observation of \mathbf{M} .

The algorithm begins with $\mathbf{R}^{(0)} = \mathbf{I}_{n_1}$. In the t -th step, the algorithm picks one coordinate (say $i \in [n_1]$) that violates

$\mu_i(\mathbf{R}^{(t-1)}\mathbf{M}) \leq \mu_i^*$, and scales the i -th diagonal entry of \mathbf{R} by a factor $\sqrt{1-\gamma}$ where $\gamma \in (0, 1)$ can be determined by line search. For the ℓ_1 hinge loss function, the best step size can be computed in a closed form (see Theorem 3).

The coordinate descent algorithm is equivalent to a series of *rank one row weightings*:

$$\mathbf{R}^{(T)}\mathbf{M} = \mathbf{W}^{(T)} \dots \mathbf{W}^{(2)}\mathbf{W}^{(1)}\mathbf{M},$$

where $\mathbf{W}^{(t)}$ is a diagonal matrix with all but one diagonal entries equal to one. Lemma 1 indicates that the leverages after rank one row weighting can be computed in a closed form. The closed-form solution makes computation and analysis much simpler; this is the reason why we use coordinate descent to solve model (8).

Given $\mu_i(\mathbf{M})$ and the desired leverage score μ_i^* , we can compute a weight γ in order that $\mu_i(\mathcal{W}(n_1, i, \gamma)\mathbf{M}) = \mu_i^*$. We show this in the following corollary, which follows directly from Lemma 1.

COROLLARY 2. *Given any $n_1 \times n_2$ matrix \mathbf{M} and any index $i \in [n_1]$. Suppose we are given $\mu_i(\mathbf{M}) \in (0, 1)$ and the desired leverage score $\mu_i^* \in [0, \mu_i(\mathbf{M})]$. We compute γ by*

$$\gamma = \frac{1 - \mu_i^*/\mu_i(\mathbf{M})}{1 - \mu_i^*}, \quad (12)$$

and scale the i -th row of \mathbf{M} by $\sqrt{1-\gamma}$. Then we have that $\mu_i(\mathcal{W}(n_1, i, \gamma)\mathbf{M}) = \mu_i^*$.

Let i be the index selected in the t -th step of the coordinate descent. Analysis shows that setting such a γ that $\mu_i(\mathbf{R}^{(t)}\mathbf{M}) = \mu_i^*$ leads to the steepest descent in the ℓ_1 hinge loss function. Therefore, if $\mu_i(\mathbf{R}^{(t-1)}\mathbf{M})$ is known to us, we can set γ according to the preceding corollary. This property is summarized in the following theorem.

THEOREM 3. *In the t -th step of the coordinate descent, suppose that the index i is selected and that $\mu_i(\mathbf{R}^{(t-1)}\mathbf{M})$ is known to us. Then scaling the i -th row by $\sqrt{1-\gamma^{(t)}}$ leads to the steepest descent in the ℓ_1 hinge loss function, where $\gamma^{(t)}$ is computed by*

$$\gamma^{(t)} = \frac{1 - \mu_i^*/\mu_i(\mathbf{R}^{(t-1)}\mathbf{M})}{1 - \mu_i^*},$$

4.3 Comparing among Different Loss Functions

Now we discuss how to set q in the the optimization problem (8). We compare ℓ_∞ , ℓ_2 , and ℓ_1 hinge loss functions and conclude that the ℓ_1 hinge loss function may be the best choice among the three for two reasons. First, optimizing the ℓ_1 hinge loss function admits provable descent in the objective function. Second, optimizing the ℓ_1 function leads to big step size, optimizing the ℓ_2 function usually leads to small step size, and optimizing the ℓ_∞ function can get stuck in some very bad configurations. The detailed discussion and empirical comparisons are presented in Appendix B. In the rest of this paper, we consider optimizing the ℓ_1 hinge loss function.

5. LEVERAGE SCORE ESTIMATION AND INEXACT ROW WEIGHTING

The essence of the proposed coordinate descent algorithm is to find a series of weights $\gamma^{(1)}, \dots, \gamma^{(T)}$ and scale the selected rows accordingly. However, to compute the weights, we need to know the leverage scores of $\mathbf{R}^{(0)}\mathbf{M}, \dots, \mathbf{R}^{(T-1)}\mathbf{M}$, at least approximately. In Section 5.1 we provide a provable approach to estimate the leverage scores. Then in Sections 5.2 and 5.3 we use the estimated leverage scores to compute the weights, and provide bounds on the true leverage scores of the matrix after weighting.

5.1 Additive-Error Bound on Leverage Scores

Suppose the $n_1 \times n_2$ rank k matrix \mathbf{M} is the ground truth, and we only have a perturbed observation of \mathbf{M} , denote $\tilde{\mathbf{M}}$. The matrix perturbation theory of [26] indicates that we can get a rough estimate of $\mathbf{U}_{\mathbf{M}} \in \mathbb{R}^{n_1 \times k}$ using the SVD of $\tilde{\mathbf{M}}$. Since the row leverage scores are the squared ℓ_2 norms of rows of $\mathbf{U}_{\mathbf{M}} \in \mathbb{R}^{n_1 \times k}$, so it is possible to approximately compute the leverage scores of \mathbf{M} through the SVD of $\tilde{\mathbf{M}}$.

Based on the matrix perturbation theory of singular vectors, [12] studied the perturbation bounds of leverage scores when the observation $\tilde{\mathbf{M}}$ is the superposition of \mathbf{M} and data noise Δ . [12] also conducted empirical studies and concluded that large leverage scores have small relative perturbation and that small leverage scores yield large relative perturbation. Although their results and analysis do not apply to the matrix completion problem, their work motivates us to estimate the large leverage scores based on the observation $\mathcal{P}_{\Omega}(\mathbf{M})$.

Under the uniform sampling model, we establish an additive-error perturbation bound for the (cross) leverage scores based on the previous work of [15, 13]. If we are given a partial observation of \mathbf{M} with $|\Omega| = \mathcal{O}(nk^2\rho^2\kappa^2(\mathbf{M}))$, Theorem 4 ensures that an estimate of the leverage scores of \mathbf{M} up to an additive-error of $\pm \frac{1}{2\rho}$ can be obtained. In this way, the rows with high leverage scores (i.e., greater than $\frac{1}{\rho}$) can be identified and then scaled. Here $\kappa(\mathbf{M}) = \sigma_1(\mathbf{M})/\sigma_k(\mathbf{M})$ is the condition number of the rank k matrix \mathbf{M} .

THEOREM 4. *Let \mathbf{M} be an $n_1 \times n_2$ matrix of rank k , $n = \max\{n_1, n_2\}$, Ω be an index set whose elements are chosen from $[n_1] \times [n_2]$ uniformly at random, and $\tilde{\mathbf{M}} = \text{Trim}(\mathcal{P}_{\Omega}(\mathbf{M}))$. When the sample complexity satisfies*

$$|\Omega| \geq C\kappa^2(\mathbf{M})nk^2\rho^2$$

for a large enough constant C , we have that with probability at least $1 - n^{-3}$ all the leverage scores and the cross leverage scores satisfy that

$$\begin{aligned} |\mu_i(\mathbf{M}) - \mu_i(\tilde{\mathbf{M}}_k)| &\leq \frac{1}{2\rho}, \\ |\mu_{ij}(\mathbf{M}) - \mu_{ij}(\tilde{\mathbf{M}}_k)| &\leq \frac{1}{2\rho}, \end{aligned}$$

for all $i, j \in \{1, \dots, n_1\}$.

In the statement of the theorem, “ $\text{Trim}(\mathcal{P}_{\Omega}(\mathbf{M}))$ ” is the operation that sets to zero all rows in $\mathcal{P}_{\Omega}(\mathbf{M})$ with degrees larger than $2|\Omega|/n_1$ and all columns in $\mathcal{P}_{\Omega}(\mathbf{M})$ with degrees larger than $2|\Omega|/n_2$. [15] show that the trim operation is necessary to ensure the bound on $\|\mathbf{M} - \tilde{\mathbf{M}}\|_2$. In practice, it is not necessary to throw away such rows and columns; if a row has degree larger than $2|\Omega|/n_1$, we can randomly hold $|\Omega|/n_1$ entries of that row.

5.2 Weighting Rows with the Medium-Scale Leverage Scores

We apply Theorem 4 to enable provable row weighting using the estimated leverage scores instead of the exact ones. Suppose the leverage scores of \mathbf{M} can be estimated within additive error $\pm \frac{1}{\rho}$. The following theorem considers a row with estimated leverage score in this region: $\hat{\mu}_i \in [\frac{1}{\rho}, 1 - \frac{1}{\rho}]$.

THEOREM 5. *Let \mathbf{M} be an $n_1 \times n_2$ rank k matrix which is unknown to us. The leverage scores of \mathbf{M} are all in $(0, 1)$. Suppose we are given the estimated leverages $\hat{\mu}_1, \dots, \hat{\mu}_n$ such that*

$$|\mu_i(\mathbf{M}) - \hat{\mu}_i| \leq \frac{1}{2\rho} \quad \text{for all } i \in [n].$$

Algorithm 2 Coordinate Descent Using Estimated Leverage Scores.

```

1: Input: an  $n_1 \times n_2$  matrix  $\tilde{\mathbf{M}}$  and a target rank  $k$ ;
2: Initialize:  $\mathbf{R}^{(0)} = \mathbf{I}_{n_1}$ ,  $\rho^{(0)} = \sqrt{\frac{|\Omega|}{Cnk^2\kappa^2(\mathbf{M})}} \gg 1$ ;
3: for  $t = 1$  to  $T$  do
4:    $\hat{\mu}_i \leftarrow$  the leverage scores of  $(\mathbf{R}^{(t-1)}\tilde{\mathbf{M}})_k$ , for all  $i \in [n_1]$ ;
5:   Pick an index such that  $\hat{\mu}_i \geq \frac{1}{\rho^{(t-1)}}$ ; Stop if there is no such  $i$ ;
6:   If  $\hat{\mu}_i \in [\frac{1}{\rho^{(t-1)}}, 1 - \frac{1}{\rho^{(t-1)}}]$ , compute  $\gamma$  according to (13);
7:   If  $\hat{\mu}_i \geq 1 - \frac{1}{\rho^{(t-1)}}$ , compute  $\gamma$  according to (14);
8:    $\mathbf{W} = \mathcal{W}(\gamma, n, i)$ ;
9:    $\mathbf{R}^{(t)} \leftarrow \mathbf{W}\mathbf{R}^{(t-1)}$ ;  $\rho^{(t)} \leftarrow \frac{\kappa(\mathbf{M})}{\kappa(\mathbf{R}^{(t)}\mathbf{M})}\rho^{(0)}$ ;
10: end for
11: Output:  $\mathbf{R} = \mathbf{R}^{(T)}$ .
```

Then for any row $i \in [n_i]$ whose estimated leverage is $\hat{\mu}_i \in [\frac{1}{\rho}, 1 - \frac{1}{\rho}]$, we let

$$\gamma = \frac{n_1 - 2k/\hat{\mu}_i}{n_1 - 2k} \quad (13)$$

and compute the diagonal matrix $\mathbf{W} = \mathcal{W}(n_1, i, \gamma)$ according to (3). The true i -th leverage score after the rank one row weighting is

$$\mu_i(\mathbf{WM}) \in \left(\frac{k}{n_1}, \frac{4k}{n_1}(1 + o(1)) \right).$$

Recall from (11) that under the uniform sampling model, the ideal leverage scores are $\mu_1 = \dots = \mu_{n_1} = \frac{k}{n_1}$. Thus weighting a row according to the theorem makes its leverage score more desirable.

5.3 Weighting Rows with the Leverage Scores Close to One

Theorem 5 indicates that when the estimated leverage score $\hat{\mu}_i \in [\frac{1}{\rho}, 1 - \frac{1}{\rho}]$, we can safely scale the i -th row by the factor $\sqrt{1 - \gamma}$, where γ is a function of $\hat{\mu}_i$ defined in (13). We plot $\hat{\mu}_i$ versus $\sqrt{1 - \gamma}$ in Figure 1.

We can see that when $\hat{\mu}_i \in [\frac{1}{\rho}, 1 - \frac{1}{\rho}]$, a small perturbation in $\hat{\mu}_i$ leads to a small change in $\sqrt{1 - \gamma}$. However, when $\hat{\mu}_i$ is close to one, a small perturbation in $\hat{\mu}_i$ leads to a big change in $\sqrt{1 - \gamma}$. Thus, taking a big step size as in (13) can be hazardous when $\hat{\mu}_i$ is close to one. When $\hat{\mu}_i > 1 - \frac{1}{\rho}$, we should set γ small and gradually and safely decrease the i -th leverage score. This strategy is described and analyzed in the following theorem.

THEOREM 6. *Let the assumptions of Theorem 5 hold except for that $\hat{\mu}_i \in (1 - \frac{1}{\rho}, 1)$. We compute γ by*

$$\gamma = \frac{\rho - \frac{1}{\hat{\mu}_i - 1/2\rho}}{\rho - 1}, \quad (14)$$

and let $\mathbf{W} = \mathcal{W}(n_1, i, \gamma)$. Then we have $\mu_i(\mathbf{WM}) \in [\frac{1}{\rho}, \mu_i(\mathbf{M})]$.

By scaling the i -th row multiple times, we can make sure that $\hat{\mu}_i$ falls in the interval $[\frac{1}{\rho}, 1 - \frac{1}{\rho}]$, and can then take a big step size according to Theorem 5.

6. PRACTICAL COORDINATE DESCENT ALGORITHM

In this section we provide a practical coordinate descent algorithm to optimize the ℓ_1 hinge loss function under the uniform sampling model without knowing the exact leverage scores. The algorithm is described in Algorithm 2. The algorithms proposed in this section do not directly apply to the non-uniform sampling model because currently we do not know how to estimate the leverage scores from non-uniformly sampled entries.

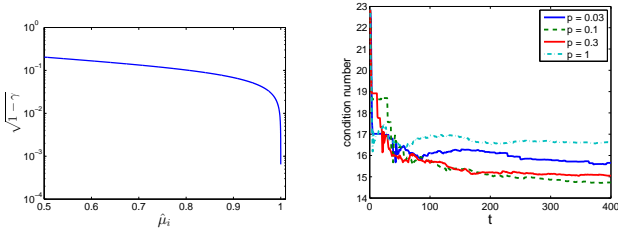


Figure 1: The plot of $\hat{\mu}_i$ versus $\sqrt{1-\gamma}$, where $\hat{\mu}_i$ and γ are condition number $\kappa(\mathbf{R}^{(t)}\mathbf{M})$ defined (13) with $k = 20$, $n_1 = 1,000$, and $\rho = 20$. **Figure 2:** The change in condition number during the rank one row weightings.

6.1 Algorithm Description

The algorithm is based on the same idea as the one in Section 4.2 except that here we use the estimated leverage scores instead of the exact ones. To optimize the ℓ_1 hinge loss function in (11), in each step the algorithm picks an index $i \in [n]$ which violates $\mu_i(\mathbf{R}\mathbf{M}) \leq \frac{k}{n_1}$. Notice that our estimated leverage scores have $\pm \frac{1}{2\rho}$ additive error, thus only the leverage score greater than $\frac{k}{n_1} + \frac{1}{2\rho}$ can be identified.

We let $\widetilde{\mathbf{M}}$ be the observation, which is $\mathcal{P}_\Omega(\mathbf{M})$ in the matrix completion problem. We let i be the selected index and $\hat{\mu}_i$ be the estimated leverage score of $\mathbf{R}^{(t-1)}\mathbf{M}$. We can compute the weight γ according to Theorems 5 and 6 and weight the i -th row by $\sqrt{1-\gamma}$. This kind of row weighting has provable bound on the true leverage scores, and thus leads to provable decrease in the objective function of (11).

6.2 Analysis

In this subsection we show that each iteration in Algorithm 2 does not increase the objective function value (11) and decrease the objective function value under some conditions.

It follows from Lemma 1 that the decrement in the i -th leverage score turns to the increments in the j -th leverage scores for all $j \neq i$. We denote the increments in the j -th ($j \neq i$) leverage scores by

$$\begin{aligned} \Delta\mu_j &:= \mu_j(\mathbf{R}^{(t)}\mathbf{M}) - \mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) \\ &= \frac{\mu_{ij}^2(\mathbf{R}^{(t-1)}\mathbf{M})}{\sum_l \mu_{il}^2(\mathbf{R}^{(t-1)}\mathbf{M})} (\mu_i(\mathbf{R}^{(t-1)}\mathbf{M}) - \mu_i(\mathbf{R}^{(t)}\mathbf{M})) \geq 0. \end{aligned} \quad (15)$$

We let \mathcal{J} be the index set

$$\mathcal{J} = \left\{ j \in [n_1] \mid j \neq i \text{ and } \mu_{ij}(\mathbf{R}^{(t-1)}\mathbf{M}) \neq 0 \text{ and } \mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) < \mu_j^* \right\}. \quad (16)$$

The following theorem ensures that the ℓ_1 hinge loss function value does not increase and strictly decrease when \mathcal{J} is nonempty.

THEOREM 7. *In each iteration of the algorithm, the following inequality holds:*

$$L_{\mathbf{M},1}(\mathbf{R}^{(t)}) \leq L_{\mathbf{M},1}(\mathbf{R}^{(t-1)}).$$

The decrement in the objective function is

$$\begin{aligned} &L_{\mathbf{M},1}(\mathbf{R}^{(t-1)}) - L_{\mathbf{M},1}(\mathbf{R}^{(t)}) \\ &= \sum_{j \in \mathcal{J}} \min \left\{ \Delta\mu_j, \mu_j^* - \mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) \right\} \geq 0. \end{aligned} \quad (17)$$

Let $i \in [n_1]$ be the index selected in the t -th iteration. Suppose that \mathcal{J} is nonempty, equivalently, at least one row (say $j \neq i$) satisfies

$$\mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) < \mu_j^* \quad \text{and} \quad \mu_{ij}(\mathbf{R}^{(t-1)}\mathbf{M}) > 0.$$

Then we have that

$$L_{\mathbf{M},1}(\mathbf{R}^{(t)}) < L_{\mathbf{M},1}(\mathbf{R}^{(t-1)}).$$

The following corollary indicates that γ is the greater the better when $\mu_i(\mathbf{R}^{(t)}\mathbf{M}) \geq \mu_i^*$ holds. Therefore, under the uniform sampling model, setting γ according to Theorem 5 leads to the greatest provable decrease in the ℓ_1 hinge loss function.

COROLLARY 8. *Let $i \in [n_1]$ be the index selected in the t -th iteration, that is, $\mu_i(\mathbf{R}^{(t-1)}\mathbf{M}) > \mu_i^*$. Since $\mu_i(\mathbf{R}^{(t)}\mathbf{M})$ decreases as γ increases, we assume that γ is small enough such that $\mu_i(\mathbf{R}^{(t)}\mathbf{M}) \geq \mu_i^*$, equivalently, assume that*

$$\gamma \leq \frac{1 - \mu_i^* / \mu_i(\mathbf{R}^{(t-1)}\mathbf{M})}{1 - \mu_i^*}.$$

Then the decrement in the ℓ_1 hinge loss function increases as γ increases.

REMARK 1. *To perform the rank one row weighting, we need to obtain an estimate of $\mu_i(\mathbf{R}^{(t-1)}\mathbf{M})$ within additive error $\pm \frac{1}{2\rho}$ for all $i \in [n_1]$ and $t \in [T]$. If we use Theorem 4 to estimate the leverage score, the parameter ρ is not constant. In fact, in the t -th iteration, the parameter ρ becomes*

$$\rho^{(t)} = \sqrt{\frac{|\Omega|}{Cnk^2\kappa^2(\mathbf{R}^{(t)}\mathbf{M})}} = \frac{\kappa(\mathbf{M})}{\kappa(\mathbf{R}^{(t)}\mathbf{M})} \rho^{(0)}.$$

Here n , k , κ , C are defined in Theorem 4, and larger ρ is more desirable. Unfortunately, how to estimate the condition numbers $\kappa(\mathbf{R}^{(t)}\mathbf{M}) = \sigma_1(\mathbf{R}^{(t)}\mathbf{M}) / \sigma_k(\mathbf{R}^{(t)}\mathbf{M})$ remains unknown, so we are unable to update $\rho^{(t)}$ in each iteration. We have to empirically fix $\rho^{(0)} = \dots = \rho^{(T-1)} = \rho$ and assume that the condition number $\kappa(\mathbf{R}^{(t)}\mathbf{M})$ does not deteriorate much during the sequence of rank one row weightings.

In practice, the assumption that the condition number $\kappa(\mathbf{R}^{(t)}\mathbf{M})$ does not deteriorate holds. In fact, in our experiments, the condition number actually gets better during the sequence of rank one row weightings. Figure 2 shows the changes of the condition number of the matrix $\mathbf{R}^{(t)}\mathbf{M}$ under the same experimental setting as that of Section 8.3.

6.3 Computational Issues

The most expensive operation in Algorithm 2 is computing the leverage score of $(\mathbf{R}^{(t-1)}\widetilde{\mathbf{M}})_k$ for all $t \in [T]$, which requires the rank k truncated SVD of $\mathbf{R}^{(t-1)}\widetilde{\mathbf{M}}$ and costs time $\mathcal{O}(Tn_1n_2k)$. The memory cost of Algorithm 2 is $\mathcal{O}(\|\widetilde{\mathbf{M}}\|_0 + n_1k)$, which is not a big challenge. We seek to reduce the time costs in the following ways.

If $\text{rank}(\widetilde{\mathbf{M}}) = r$ and $k \leq r < n_2$, we can reduce the time cost to $\mathcal{O}(n_1n_2r + Tn_1rk)$ in the following way. We first compute the condensed SVD of $\widetilde{\mathbf{M}}$ to obtain its column bases $\mathbf{U}_{\widetilde{\mathbf{M}}} \in \mathbb{R}^{n_1 \times r}$ in time $\mathcal{O}(n_1n_2r)$. Exploiting the fact that the leverage scores of $\mathbf{R}\widetilde{\mathbf{M}}$ and $\mathbf{R}\mathbf{U}_{\widetilde{\mathbf{M}}}$ are the same for any nonsingular matrix \mathbf{R} (see Theorem 11 in the appendix), we replace $\widetilde{\mathbf{M}} \in \mathbb{R}^{n_1 \times n_2}$ in Algorithm 2 by the smaller matrix $\mathbf{U}_{\widetilde{\mathbf{M}}} \in \mathbb{R}^{n_1 \times r}$. Then in each iteration it cost only $\mathcal{O}(n_1rk)$ time to compute the leverage scores of $\mathbf{R}^{(t-1)}\mathbf{U}_{\widetilde{\mathbf{M}}}$.

Algorithm 3 The Weighting—Completion Algorithm.

```

1: Input: the partially observed matrix  $\mathcal{P}_\Omega(\mathbf{M})$ , a target rank  $k$ .
2:  $\widetilde{\mathbf{M}}^{(0)} \leftarrow \text{Trim } \mathcal{P}_\Omega(\mathbf{M})$  according to [15];
3: for  $s = 1, 2, \dots$  do
4:    $\mathbf{R} \leftarrow$  Algorithm 2 taking  $\widetilde{\mathbf{M}}^{(s)}$  and  $k$  as input;
5:    $\mathbf{C} \leftarrow$  Algorithm 2 taking  $(\widetilde{\mathbf{M}}^{(s)})^T$  and  $k$  as input;
6:    $\widetilde{\mathbf{M}}^{(s)} \leftarrow$  weighted matrix completion taking  $\mathcal{P}_\Omega(\mathbf{M})$ ,  $\mathbf{R}$ ,  $\mathbf{C}$  as input;
7: end for
8: Output:  $\widetilde{\mathbf{M}}^{(s)}$ .

```

If $\text{rank}(\widetilde{\mathbf{M}})$ is not much smaller than n_2 , the above approach does not help accelerating computation. Since $(\mathbf{R}^{(t-1)}\widetilde{\mathbf{M}})_k$ is merely a rough approximation of $\mathbf{R}^{(t-1)}\mathbf{M}$, it is unnecessary to compute the exact rank k SVD of $(\mathbf{R}^{(t-1)}\widetilde{\mathbf{M}})_k$. Instead, we propose to compute the rank k SVD of $\mathbf{R}^{(t-1)}\widetilde{\mathbf{M}}$ approximately using the matrix sketching techniques [27]. For example, when $\widetilde{\mathbf{M}} = \mathcal{P}_\Omega(\mathbf{M})$, which is a highly sparse matrix, the rank k SVD of $\widetilde{\mathbf{M}}$ can be computed in time $\mathcal{O}(|\Omega| + \text{poly}(k))$ by the sparse matrix embedding method [6]. In this way, the total time cost drops to $\mathcal{O}(T|\Omega| + T \cdot \text{poly}(k))$.

7. THE WEIGHTING—COMPLETION ALGORITHM

In practice, we find that the algorithm can gradually make all the leverage score below $\frac{1}{\rho}$. However, the algorithm can do nothing to the leverage scores between $\frac{k}{n_1}$ and $\frac{1}{\rho}$ because a very small leverage score may appear large due to the additive error perturbation. Therefore, the row matrix coherence after the row weighting is $\frac{n_1}{\rho k}$, ideally. If we want to attain an even lower matrix coherence, we must acquire more accurate estimates of the leverage scores. For this purpose, we propose a heuristic for obtaining better estimates of the leverage scores.

We can first use $\mathcal{P}_\Omega(\mathbf{M})$ to estimate the row and column leverage scores and then perform matrix completion using the weighted matrix completion model (5) to obtain \mathbf{L}^* . Empirically, compared with $\mathcal{P}_\Omega(\mathbf{M})$, the leverage scores of \mathbf{L}^* better approximates those of \mathbf{M} , and we can thus obtain better estimation of the leverage scores of \mathbf{M} based on \mathbf{L}^* . We propose to perform once more row and column weighting using \mathbf{L}^* and k as the input of Algorithm 2. We can also repeat this weighting—completion procedure multiple times to attain better results. This weighting—completion procedure is described in Algorithm 3.

8. EXPERIMENTS

We conduct experiments to evaluate the coordinate descent algorithm and the weighted matrix completion. The data and algorithms are all for the uniform sampling model.

8.1 Datasets

To demonstrate the effectiveness of our approach, we generate an $n_1 \times n_2$ low-rank matrix $\mathbf{L}_0 = \mathbf{U}\mathbf{V}^T$ with high row and column coherences. We generate $\mathbf{U} \in \mathbb{R}^{n_1 \times k}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times k}$ with each row sampled from the multivariate t distribution with 2 degrees of freedom and the covariance matrix $\mathbf{\Lambda} \in \mathbb{R}^{k \times k}$ whose (i, j) -th entry is $\Lambda_{ij} = 2 \times 0.5^{|i-j|}$. We generate an index set Ω by sampling each element of $[n_1] \times [n_2]$ with probability p . Thus, the expected number of observed entries is $\mathbb{E}|\Omega| = pn_1n_2$. The $n_1 \times n_2$ matrix $\mathcal{P}_\Omega(\mathbf{M})$ is our observation, where $\mathbf{M} = \mathbf{L}_0 + \mathbf{S}_0$ and \mathbf{S}_0 captures the noise.

Our weighted approach does not achieve markedly advantage over the unweighted on the collaborative filtering data. The collaborative filtering data are highly unbalanced, and thus the uniform sampling model considered in this paper is violated. Our proposed algorithms in Section 6 does not apply to the general non-uniform sampling model because the leverage scores cannot be accurately estimated using our approach. Our tentative experiments on synthetic data shows that naively estimating the leverage scores in the same way as the uniform sampling model does not work well: the weighted approach does not have noticeable advantage over the unweighted approach unless a large portion of entries were observed, e.g. 20% entries were observed. However, in the collaborative filtering problems, commonly less than 1% entries are observed.

8.2 Compared Methods

We use model (6) under different settings of the weight matrices \mathbf{R} and \mathbf{C} . If k (the rank of the underlying low-rank matrix) is unknown, we treat it as a parameter and tune it.

- **Unweighted.** Set $\mathbf{R} = \mathbf{I}_{n_1}$ and $\mathbf{C} = \mathbf{I}_{n_2}$.
- **Type 1.** Take $\mathcal{P}_\Omega(\mathbf{M})$ and k as the input of Algorithm 2 to compute \mathbf{R} and \mathbf{C} , and solve model (6) to obtain the solution \mathbf{L}^* . That is, run Algorithm 3 with only one repeat.
- **Type 2.** Let \mathbf{L}^* be computed by Type 1. Take $\widetilde{\mathbf{M}} = \mathbf{L}^*$ and k as the input of Algorithm 2 to re-compute \mathbf{R} and \mathbf{C} . That is, run Algorithm 3 for two repeats. **Type 3, Type 4, ...**, are similarly defined.

For each data matrix and each method, we tune the parameter λ to the best and report the matrix completion error

$$\text{Error} = \|\mathbf{L}^* - \mathbf{L}_0\|_F / \|\mathbf{L}_0\|_F, \quad (18)$$

where \mathbf{L}_0 denotes the ground truth, i.e., the low-rank component of \mathbf{M} .

8.3 Effects on Leverage Scores

In this subsection we test how the leverage scores changes in each step of the coordinate descent algorithm 2. Let T, t, ρ be defined in Algorithm 2, and set $\rho = 20\sqrt{p}$ and the total iterative number of coordinate descent to be $T = k^2$. We generate the synthetic data matrix \mathbf{L}_0 described in Section 8.1 with different settings of n_1, n_2, k , and p , and let $\mathbf{M} = \mathbf{L}_0$ without adding noise.

We first test the coordinate descent algorithm under different $p = |\Omega|/n_1n_2$. We fix $n_1 = 2,000, n_2 = 1,000, k = 20$ to generate the low rank matrix \mathbf{M} , and vary the number of samples by setting $p = 0.03, 0.1, 0.3, 1$ respectively to obtain Ω . We take $\mathcal{P}_\Omega(\mathbf{M})$ and k as the input of Algorithm 2, and report in Figure 3 the row coherence $\mu(\mathbf{R}^{(t)}\mathbf{M})$ and the ℓ_1 hinge loss function value $L_{\mathbf{M},1}(\mathbf{R}^{(t)})$ in each step of the coordinate descent algorithm, respectively.

In Figure 3 the plot of the ℓ_1 hinge loss function indicates that the leverage scores become more uniform after each rank one row weighting using Algorithm 2. Theorem 4 indicates that larger p results in better estimation of the leverage scores, and consequently the row weighting can be better performed. The experimental results verify our intuition.

We test the weighting—completion algorithm (Algorithm 3) to see whether alternating between weighting and completion helps making the leverage score more uniform. We fix $n_1 = 2,000, n_2 = 1,000, k = 20, p = \frac{|\Omega|}{n_1n_2} = 0.1$. We run the weighting—completion procedure in Algorithm 3 for four iterations and plot t against the the matrix coherence or the ℓ_1 hinge loss function $L_{\mathbf{M},1}(\mathbf{R}^{(t)})$ in Figure 4.

Figure 4 clearly shows that performing the weighting—completion procedure multiple rounds results in much lower matrix coherence

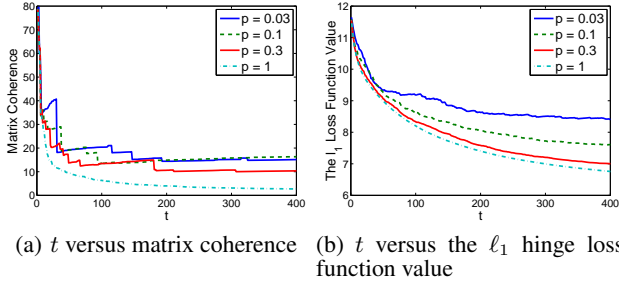


Figure 3: Adjusting leverage scores by row weighting under different number of samples.

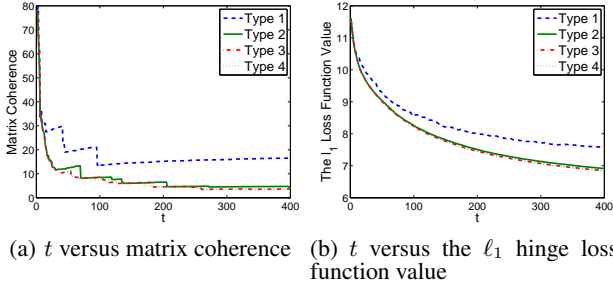


Figure 4: The effects of the weighting—completion algorithm (Algorithm 3).

and the ℓ_1 hinge loss function value than performing the procedure only one round. The results suggest that running the weighting—completion procedure for two rounds is a good choice.

8.4 Matrix Completion Accuracy

In this set of experiments, we use the synthetic data described in Section 8.1 to generate \mathbf{L}_0 and add i.i.d. Gaussian noise $\mathcal{N}(1, \sigma^2)$ to 50% entries of \mathbf{L}_0 to obtain \mathbf{M} . We set $n_1 = 2,000$, $n_2 = 1,000$, $k = 20$, and vary p and σ . We set $p = 0.05, 0.1$, or 0.2 , and vary σ from 0 to 40. For each data matrix, each p , σ , and each method, we tune the parameter λ to attain the lowest matrix completion error (defined in (18)). We plot in Figure 5 the noise intensity $\frac{\|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L}_0)\|_F}{\|\mathcal{P}_\Omega(\mathbf{M})\|_F}$ against the matrix completion error.

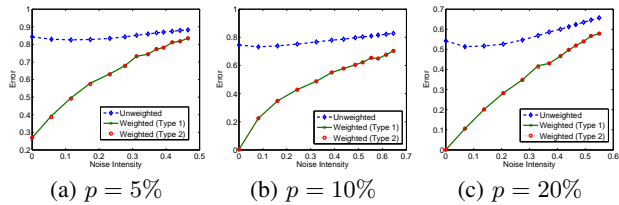


Figure 5: Experiments on noisy matrix completion.

The experimental results show that when the underlying low-rank matrix is coherent, the standard nuclear norm minimization method fails even on the noise-free data with a large number of samples, i.e., $p = 20\%$ entries observed. In comparison, our weighted method strictly succeeds on the noise-free data with

$p = 10\%$ or 20% , and our method achieves high accuracy even under heavy noise.

9. EXTENSION TO THE ROBUST PRINCIPAL COMPONENT ANALYSIS

Our proposed row/column weighting method can also be applied to low-rank matrix recovery problems other than matrix completion. The robust principal component analysis (RPCA) is a well-known low-rank plus sparse matrix recovery model, but it requires the matrix coherence to be small in order to recover the underlying low-rank and sparse matrices. In this section we apply our method to RPCA and call the obtained method the weighted RPCA. We show that highly coherent low-rank matrices can also be recovered by the weighted RPCA.

9.1 The Standard RPCA Method

In some real-world applications the observation $\mathbf{D} \in \mathbb{R}^{n_1 \times n_2}$ is the superposition of a low-rank matrix \mathbf{L}_0 and a sparse matrix \mathbf{S}_0 . It is useful to recover the low-rank matrix and the sparse matrix from the observation. For example, in the video surveillance problem, by vectorizing each frame of a video surveillance sequence and stacking the obtained vectors, the obtained matrix is the sum of the low-rank background and the sparse foreground.

The robust principal component analysis (RPCA) [2] is perhaps the most effective tool for such a matrix recovery task. The RPCA model is defined as follows:

$$\min_{\mathbf{L}, \mathbf{S}} \|\mathbf{S}\|_1 + \lambda \|\mathbf{L}\|_*; \quad \text{s.t. } \mathbf{L} + \mathbf{S} = \mathbf{D}. \quad (19)$$

Let $\tau = \max \{ \mu(\mathbf{L}_0), \nu(\mathbf{L}_0), \frac{n_1 n_2}{k} \|\mathbf{U}_{\mathbf{L}_0, k} \mathbf{V}_{\mathbf{L}_0, k}^T\|_\infty^2 \}$ and $k = \text{rank}(\mathbf{L}_0)$. Suppose $n_1 \geq n_2$. It was shown in [2] that RPCA exactly recovers \mathbf{L}_0 and \mathbf{S}_0 provided that $k \leq c_l(n_1 + n_2)\tau^{-1} \log^{-2} n_1$ and $\|\mathbf{S}_0\|_0 \leq c_s n_1 n_2$ for some constants c_l and c_s . Therefore, the exact recovery is guaranteed only when the matrix coherence parameter τ is small.

9.2 Weighted RPCA

When the underlying low-rank matrix \mathbf{L}_0 is coherent, the standard RPCA method can easily fail. To remedy this problem, we propose the use of weighted RPCA to recover coherent low-rank matrix.

Let \mathbf{R} and \mathbf{C} be some diagonal matrices. We can weight \mathbf{D} by

$$\underbrace{\mathbf{R} \mathbf{D} \mathbf{C}}_{\hat{\mathbf{D}}} = \underbrace{\mathbf{R} \mathbf{L}_0 \mathbf{C}}_{\hat{\mathbf{L}}_0} + \underbrace{\mathbf{R} \mathbf{S}_0 \mathbf{C}}_{\hat{\mathbf{S}}_0}.$$

It is obvious that $\text{rank}(\hat{\mathbf{L}}_0) = \text{rank}(\mathbf{L}_0) = k$ and $\|\hat{\mathbf{S}}_0\|_0 = \|\mathbf{S}_0\|_0$; thus $\hat{\mathbf{D}}$ is the sum of the low-rank matrix $\hat{\mathbf{L}}_0$ and the sparse matrix $\hat{\mathbf{S}}_0$. From the analysis of [2] we know that the matrix recovery performance can be improved if the matrix coherence parameters of $\hat{\mathbf{L}}_0$ is lower than those of \mathbf{L}_0 . We therefore propose to use Algorithm 2 to compute \mathbf{R} and \mathbf{C} and do row and column weighting before performing RPCA. Our proposal is as follows:

1. Compute \mathbf{R} and \mathbf{C} using Algorithm 2 which takes \mathbf{D} and k as inputs.
2. Take $\hat{\mathbf{D}} = \mathbf{R} \mathbf{D} \mathbf{C}$ instead of \mathbf{D} as the input of RPCA and obtain the solution $\hat{\mathbf{L}}^*, \hat{\mathbf{S}}^*$.
3. Output $\mathbf{L}^* = \mathbf{R}^{-1} \hat{\mathbf{L}}^* \mathbf{C}^{-1}$ and $\mathbf{S}^* = \mathbf{R}^{-1} \hat{\mathbf{S}}^* \mathbf{C}^{-1}$.

The experiments in Section 9.3 shows that the performance of row weighting deteriorates as $\|\mathbf{S}_0\|_0$ or $\|\mathbf{S}_0\|_\infty$ increases. In the same spirit with the weighting—completion algorithm in Section 7, we propose to use the recovered matrix \mathbf{L}^* instead of \mathbf{D} to compute

the weight matrices \mathbf{R} and \mathbf{C} . In these experiments we use two methods to compute the weight matrices \mathbf{R} and \mathbf{C} :

- **Type 1.** Use \mathbf{D} and k as the input of Algorithm 2 to compute \mathbf{R} and \mathbf{C} and perform the weighted RPCA to recover \mathbf{L}^* .
- **Type 2.** Use Type 1 to recover \mathbf{L}^* , and then use \mathbf{L}^* and k as the input of Algorithm 2 to obtain $\hat{\mathbf{R}}$ and $\hat{\mathbf{C}}$. Finally run the weighted RPCA once more using the new weight matrices $\hat{\mathbf{R}}$ and $\hat{\mathbf{C}}$.

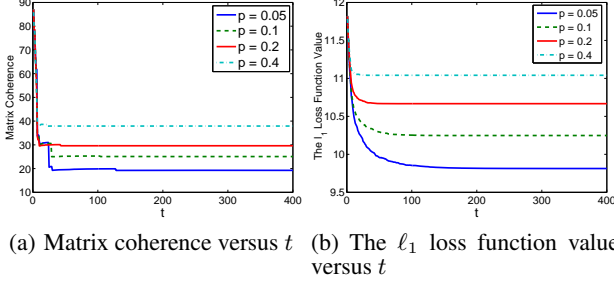


Figure 6: The effects of row weighting under varying $\|\mathbf{S}_0\|_0$.

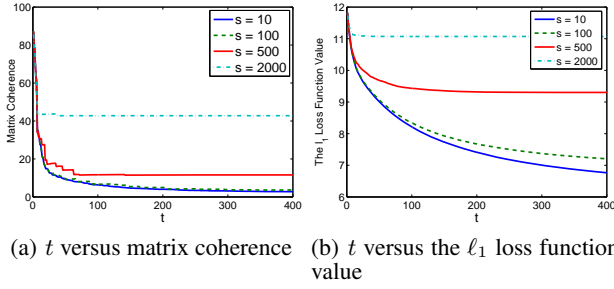


Figure 7: The effects of row weighting under varying $\|\mathbf{S}_0\|_\infty$.

9.3 Effects on Leverage Scores

We set $n_1 = 2,000$, $n_2 = 1,000$, $k = 20$ and use the data model defined in Section 8.1 to generate the low-rank matrix \mathbf{L}_0 . The sparse matrix \mathbf{S}_0 is generated independently by uniformly setting entries to be s with probability $p/2$, $-s$ with probability $p/2$, and zero with probability $1 - p$. That is $\|\mathbf{S}\|_0 = pn_1n_2$ and $\|\mathbf{S}\|_\infty = s$. We use $\mathbf{D} = \mathbf{L}_0 + \mathbf{S}_0$ as the observation.

In the first set of experiments, we illustrate the effect of Algorithm 2, which takes \mathbf{D} and k as inputs, with varying \mathbf{S}_0 . We fix $s = 1,000$ and vary p , and plot the results in Figure 6. We then fix $p = 0.1$ and vary s , and plot the results in Figure 7. The results clearly indicate that under the RPCA setting, our row weighting algorithm makes the leverage scores more uniform and the matrix coherence much lower.

9.4 Matrix Recovery Accuracy

In the second set of experiments, we compare the weighted and unweighted RPCA in terms of matrix recovery accuracy. The relative error defined in (18) is used to evaluate the recovery accuracy.

We generate the synthetic data \mathbf{L}_0 in the same way as in the previous subsection. As for the sparse matrix \mathbf{S}_0 , we first fix $s =$

1,000 and vary p , and report the relative error in Figure 8(a). We then fix $p = 0.2$ and vary s , and report the relative error in Figure 8(b).

Since the low-rank matrix \mathbf{L}_0 is highly coherent, the standard RPCA fails even if $\|\mathbf{S}_0\|_0$ and $\|\mathbf{S}\|_\infty$ are both very small. Since we use $\mathbf{D} = \mathbf{L}_0 + \mathbf{S}_0$ to compute the weight matrices \mathbf{R} and \mathbf{C} , the error in the leverage score estimation should be lower if \mathbf{D} is closer to \mathbf{L}_0 . The Type 1 weighted RPCA achieves much better performance when $\|\mathbf{S}_0\|_0$ and $\|\mathbf{S}\|_\infty$ are reasonably small, which is in accordance with our expectation. We can also see that the Type 2 weighted RPCA achieves the highest accuracy. The computational cost of the Type 2 method is only twice as much as the Type 1 method, but the accuracy is much higher than that of Type 1.

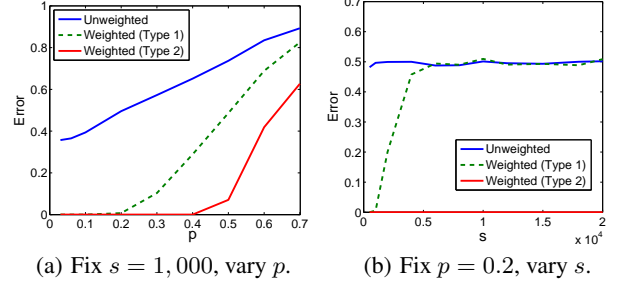


Figure 8: The matrix recovery errors.

10. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a row/column weighting method for adjusting the leverage scores such that low-rank matrix recovery can be better performed. We have established an optimization model that describes the discrepancy between the leverage scores and desired leverage scores. We have also devised a coordinate descent algorithm for solving the model without knowing the true leverage scores. We have proved that this algorithm does not increase the objective function value and that it decreases the objective function value under certain conditions. We have applied our matrix weighting method to the matrix completion problem such that coherent low-rank matrices can be completed more accurately via the weighted nuclear norm minimization formulation. Moreover, we have applied our matrix weighting method to the robust principal component analysis problem, so that highly coherent low-rank matrices can be recovered from noisy observations more accurately.

We should point out that currently our proposed method applies only to the uniform sampling model; that is, all matrix entries are observed with the same probability. Our row weighting method critically relies on the quality of leverage score estimation. For the non-uniform sampling model, it is not clear how to accurately estimate the leverage scores from partial observation. For the non-uniform sampling model, however, if we have a method that can accurately estimate the leverage scores, the same coordinate descent algorithm can be used to find the weight matrix \mathbf{R} by solving (8). Considering that most real-world applications obey non-uniform sampling settings such as the power law distribution, it is very useful to find a way to estimating leverage scores from non-uniformly sampled entries, with which our row weighting approach can demonstrate its power in the real-world applications.

11. REFERENCES

- [1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- [2] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.
- [3] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.
- [4] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.
- [5] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Coherent matrix completion. In *Proceedings of The 31st International Conference on Machine Learning (ICML)*, 2014.
- [6] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *Annual ACM Symposium on theory of computing (STOC)*. ACM, 2013.
- [7] M. B. Cohen, Y. T. Lee, C. Musco, C. Musco, R. Peng, and A. Sidford. Uniform sampling for matrix approximation. *arXiv preprint arXiv:1408.5099*, 2014.
- [8] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. The Yahoo! music dataset and KDD-Cup’11. In *KDD Cup*, pages 8–18, 2012.
- [9] R. Foygel, O. Shamir, N. Srebro, and R. Salakhutdinov. Learning with the weighted trace-norm under arbitrary sampling distributions. In J. Shawe-taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*. 2011.
- [10] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–77. ACM, 2011.
- [11] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2117–2130, 2013.
- [12] I. Ipsen and T. Wentworth. Sensitivity of leverage scores. *arXiv preprint arXiv:1402.0957*, 2014.
- [13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.
- [14] H. Ji, C. Liu, Z. Shen, and Y. Xu. Robust video denoising using low rank matrix completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [15] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.
- [16] A. Krishnamurthy and A. Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [17] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.
- [18] B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.
- [19] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.
- [20] J. D. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.
- [21] R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems*, pages 2056–2064, 2010.
- [22] A. SIGKDD. Netflix. In *Proceedings of kdd cup and workshop*, 2007.
- [23] A. M.-C. So and Y. Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2-3):367–384, 2007.
- [24] N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.
- [25] S. Wang and Z. Zhang. Colorization by matrix completion. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2012.
- [26] P.-rA. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.
- [27] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.
- [28] H.-F. Yu, C.-J. Hsieh, S. Si, and I. S. Dhillon. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. In *ICDM*, pages 765–774, 2012.
- [29] H. Yun, H.-F. Yu, C.-J. Hsieh, S. Vishwanathan, and I. Dhillon. NOMAD: Non-locking, stOchastic Multi-machine algorithm for Asynchronous and Decentralized matrix completion. In *International Conference on Very Large Data Bases (VLDB)*, 2014.
- [30] X. Zhang, D. Schuurmans, and Y. Yu. Accelerated training for matrix-norm regularization: A boosting approach. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*. 2012.
- [31] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Algorithmic Aspects in Information and Management*, pages 337–348. Springer, 2008.

APPENDIX

A. ADMM FOR WEIGHTED NUCLEAR NORM MINIMIZATION

The regularized weighted nuclear norm minimization formulation (6) can be equivalently converted to (20) by adding the slack variable \mathbf{X} :

$$\begin{aligned} \min_{\mathbf{L}, \mathbf{X}} \quad & \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L})\|_F^2 + \lambda \|\mathbf{X}\|_*, \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{RLC}. \end{aligned} \quad (20)$$

The augmented Lagrange function of this problem is

$$\begin{aligned} f_\gamma(\mathbf{L}, \mathbf{X}, \mathbf{Y}) &= \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L})\|_F^2 + \lambda \|\mathbf{X}\|_* \\ &\quad + \langle \mathbf{Y}, \mathbf{RLC} - \mathbf{X} \rangle + \frac{\gamma}{2} \|\mathbf{RLC} - \mathbf{X}\|_F^2 \\ &= \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L})\|_F^2 + \lambda \|\mathbf{X}\|_* \\ &\quad + \frac{\gamma}{2} \|\gamma^{-1} \mathbf{Y} + \mathbf{RLC} - \mathbf{X}\|_F^2 - \frac{\gamma}{2} \|\gamma^{-1} \mathbf{Y}\|_F^2. \end{aligned}$$

We alternately minimize f_γ w.r.t. \mathbf{L} and \mathbf{X} and maximize f_γ w.r.t. \mathbf{Y} .

The terms in f_γ containing $\mathcal{P}_\Omega(\mathbf{L})$ is

$$f_\gamma(\mathcal{P}_\Omega(\mathbf{L})) = \frac{1}{2} \|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L})\|_F^2 + \frac{\gamma}{2} \left\| \mathcal{P}_\Omega\left(\frac{\mathbf{Y}}{\gamma} - \mathbf{X} + \mathbf{RLC}\right) \right\|_F^2,$$

and the minimizer is

$$L_{ij}^* = (1 + \gamma R_{ii}^2 C_{jj}^2)^{-1} (M_{ij} + \gamma R_{ii} C_{jj} X_{ij} - R_{ii} C_{jj} Y_{ij}) \quad (21)$$

for all $(i, j) \in \Omega$.

Let $\mathcal{P}_\Omega^\perp(\mathbf{L})$ be the $n_1 \times n_2$ matrix whose the (i, j) -th entry is L_{ij} if $(i, j) \notin \Omega$ and is zero otherwise. The terms in f_γ containing $\mathcal{P}_\Omega^\perp(\mathbf{L})$ is

$$f_\gamma(\mathcal{P}_\Omega^\perp(\mathbf{L})) = \frac{\gamma}{2} \|\mathcal{P}_\Omega^\perp(\gamma^{-1} \mathbf{Y} - \mathbf{X} + \mathbf{RLC})\|_F^2,$$

and the minimizer is

$$L_{ij}^* = R_{ii}^{-1} C_{jj}^{-1} (X_{ij} - \gamma^{-1} Y_{ij}) \quad (22)$$

for all $(i, j) \notin \Omega$.

The terms in f_γ containing \mathbf{X} is

$$f_\gamma(\mathbf{X}) = \frac{\gamma}{2} \|\gamma^{-1} \mathbf{Y} + \mathbf{RLC} - \mathbf{X}\|_F^2 + \lambda \|\mathbf{X}\|_*,$$

and the minimizer is

$$\mathbf{X}^* = \mathbf{U}[\Sigma - \gamma^{-1} \lambda \mathbf{I}]_+ \mathbf{V}^T, \quad (23)$$

where the (i, j) of $[\mathbf{M}]_+$ is $\max\{A_{ij}, 0\}$ and

$$[\mathbf{U}, \Sigma, \mathbf{V}] \leftarrow \text{SVD}(\gamma^{-1} \mathbf{Y} + \mathbf{RLC}).$$

We can update \mathbf{L} by (21) and (22), update \mathbf{X} by (23), and update the dual variable \mathbf{Y} by gradient ascent: $\mathbf{Y} \leftarrow \mathbf{Y} + \gamma(\mathbf{RLC} - \mathbf{X})$. The whole procedure is described in Algorithm 1.

B. COMPARING AMONG DIFFERENT LOSS FUNCTIONS

We first give an example to show the advantage of optimizing the ℓ_1 hinge loss function.

In this example, if we optimize the ℓ_∞ hinge loss function by coordinate descent, scaling any row by any factor in the region $(0, 1)$ does not decrease the ℓ_∞ hinge loss function value. That is, the coordinate descent algorithm stops in this configuration.

In comparison, in this example, if we optimize the ℓ_1 hinge loss function by coordinate descent, the 1st or 2nd row will be scaled by a factor $\gamma \in (0, 1)$, and the ℓ_1 hinge loss function value will decrease in certain conditions (see Theorem 7). In this way, the coordinate descent proceeds without being stuck in the configuration in the example.

EXAMPLE 1. Let μ_1, \dots, μ_{n_1} be the row leverage scores of a rank k matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, and the ideal leverage scores are $\mu_1^* = \dots = \mu_{n-1}^* = \frac{k}{n_1}$. When $\mu_1 = \mu_2 > \frac{k}{n_1} \geq \mu_3, \dots, \mu_{n-1}$, weighting any single row by any factor $\gamma \in (0, 1)$ cannot decrease the ℓ_∞ hinge loss function value.

PROOF. If we scale the 1st row by any factor $\gamma \in (0, 1)$, the leverage score of the 2nd row will be nondecreasing (by Lemma 1). Thus the ℓ_∞ hinge loss function value does not decrease.

The same happens if we scale the 2nd row.

If we scale the i -th row for $i > 2$, the leverage scores of the 1st and 2nd rows will be nondecreasing (by Lemma 1). Thus the ℓ_∞ hinge loss function value does not decrease. \square

Here we use a toy dataset to compare among the ℓ_1 , ℓ_2 , and ℓ_∞ hinge loss functions. We generate a synthetic data matrix \mathbf{M} according to Section 8.1 by setting $n_1 = n_2 = 100$ and $k = 5$. We assume the exact leverage scores are available to us and use coordinate descent algorithm to optimize the loss functions. We use three kinds of step sizes: the ℓ_1 , ℓ_2 , and ℓ_∞ step sizes; “the ℓ_q step size” means the step size computed by line search that decrease the ℓ_q hinge loss function value. We plot in Figure 9 the loss function values versus the number of coordinate descents. The ℓ_1 step size always leads to the fastest convergence; the ℓ_1 step size does not only decrease the ℓ_1 function value fast, but also decrease the ℓ_2 and ℓ_∞ very fast. The ℓ_2 step size also works well, but it is much smaller and results in slow convergence. The experiment shows that directly minimizing the ℓ_∞ hinge loss function is problematic because it can get stuck in some bad local minimum configurations without making progress, which corroborates the analysis in Example 1.

C. PROOF OF THE THEOREMS

D. KEY LEMMAS

LEMMA 9. Let i be the index selected in the t -th step of the coordinate descent algorithm. Suppose the following equality holds:

$$\mu_i(\mathbf{R}^{(t-1)})\mathbf{M} > \mu_i(\mathbf{R}^{(t)})\mathbf{M} \geq \mu_i^*$$

Let $\Delta\mu_j$ be defined in (15) and the set \mathcal{J} be defined in (16). Then the decrement in the ℓ_1 hinge loss function is

$$\begin{aligned} & L_{\mathbf{M},1}(\mathbf{R}^{(t-1)}) - L_{\mathbf{M},1}(\mathbf{R}^{(t)}) \\ &= \sum_{j \in \mathcal{J}} \min \left\{ \Delta\mu_j, \mu_j^* - \mu_j(\mathbf{R}^{(t-1)})\mathbf{M} \right\} \geq 0, \end{aligned}$$

PROOF. Let i be the selected index in the t -th step. For convenience, we write the definition of the ℓ_1 hinge loss function here:

$$L_{\mathbf{M},1}(\mathbf{R}^{(t)}) := \sum_{l=1}^{n_1} \max \left\{ \mu_l(\mathbf{R}^{(t)})\mathbf{M} - \mu_l^*, 0 \right\}.$$

Obviously, the decrease in the i -th leverage score leads to the decrease in the loss function value, and the increase in the j -th ($j \neq i$) leverage scores may increase the loss function value.

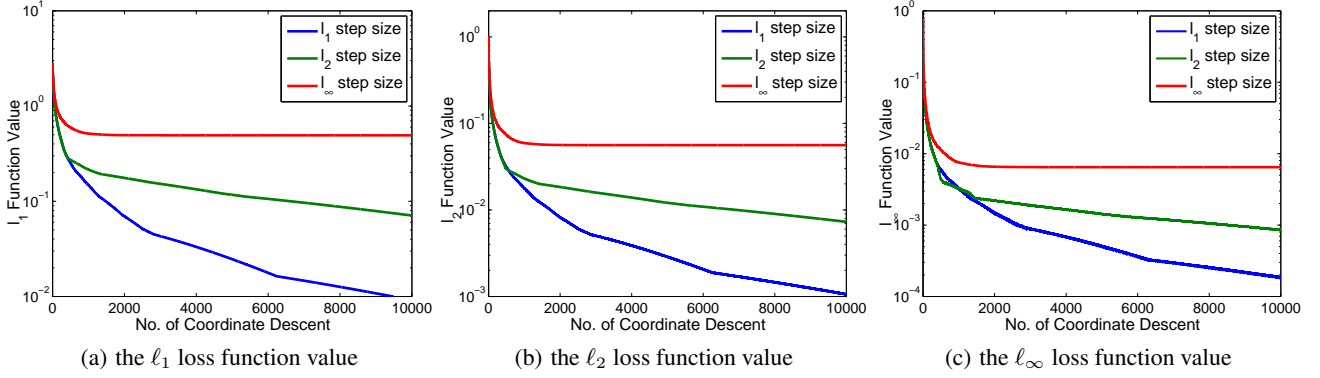


Figure 9: The comparisons among the loss functions.

The decrease in the i -th leverage score contributes to a decrease in the ℓ_1 hinge loss function:

$$\delta_1 := \mu_i(\mathbf{R}^{(t-1)}\mathbf{M}) - \mu_i(\mathbf{R}^{(t)}\mathbf{M}) > 0.$$

From Lemma 1 we know that $\mu_j(\mathbf{R}^{(t)}\mathbf{M}) \geq \mu_j(\mathbf{R}^{(t-1)}\mathbf{M})$ for all $j \neq i$. The increase in the leverage scores $\{\mu_j\}_{j \neq i}$ contributes to a increase in the ℓ_1 hinge loss function:

$$\delta_2 := \sum_{j \in \mathcal{J}_1} \Delta\mu_j + \sum_{j \in \mathcal{J}_2} \max\{\mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) + \Delta\mu_j - \mu_j^*, 0\} \geq 0$$

where $\mathcal{J}_1 = \{j \neq i \mid \mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) \geq \mu_j^*\}$ and $\mathcal{J}_2 = \{j \neq i \mid \mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) < \mu_j^*\}$. Thus the decrease the the ℓ_1 hinge loss function is

$$\begin{aligned} L_{\mathbf{M},1}(\mathbf{R}^{(t-1)}) - L_{\mathbf{M},1}(\mathbf{R}^{(t)}) &= \delta_1 - \delta_2 \\ &= \sum_{j \in \mathcal{J}_1} \Delta\mu_j + \sum_{j \in \mathcal{J}_2} \Delta\mu_j - \delta_2 \\ &= \sum_{j \in \mathcal{J}_2} \left[\Delta\mu_j - \max\{\mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) + \Delta\mu_j - \mu_j^*, 0\} \right] \\ &= \sum_{j \in \mathcal{J}_2} \min\{\Delta\mu_j, \mu_j^* - \mu_j(\mathbf{R}^{(t-1)}\mathbf{M})\} \\ &= \sum_{j \in \mathcal{J}} \min\{\Delta\mu_j, \mu_j^* - \mu_j(\mathbf{R}^{(t-1)}\mathbf{M})\}. \end{aligned}$$

Here the second equality follows from the definition of $\Delta\mu_j$ that $\delta_1 = \sum_{j \neq i} \Delta\mu_j$. The last equality follows from that $\Delta\mu_j = 0$ if $\mu_{ij}(\mathbf{R}^{(t-1)}\mathbf{M}) = 0$. \square

LEMMA 10. Let i be the index selected in the t -th step of the coordinate descent algorithm, and the i -th row of $\mathbf{R}^{(t-1)}\mathbf{M}$ is scaled by $\sqrt{1-\gamma} \in (0, 1)$ to obtain $\mathbf{R}^{(t)}\mathbf{M}$. Suppose $\mu_i(\mathbf{R}^{(t-1)}\mathbf{M}) > \mu_i^*$. Then we have

1. The leverage score $\mu_i(\mathbf{R}^{(t)}\mathbf{M})$ decrease as γ increases.
2. When $\gamma \leq \frac{1-\mu_i^*/\mu_i(\mathbf{R}^{(t-1)}\mathbf{M})}{1-\mu_i^*}$, or equivalently, $\mu_i(\mathbf{R}^{(t)}\mathbf{M}) \geq \mu_i^*$, the decrement in the ℓ_1 hinge loss function (defined in Lemma 9) increases as γ increases.
3. When $\gamma > \frac{1-\mu_i^*/\mu_i(\mathbf{R}^{(t-1)}\mathbf{M})}{1-\mu_i^*}$, or equivalently, $\mu_i(\mathbf{R}^{(t)}\mathbf{M}) < \mu_i^*$, the decrement in the ℓ_1 hinge loss function does not increase as γ increases.

PROOF. The first property follows directly from Lemma 1.

Second, from the definition of $\Delta\mu_j$ in (15), we know that $\Delta\mu_j$ increases as $\mu_i(\mathbf{R}^{(t)}\mathbf{M})$ decreases. When $\mu_i(\mathbf{R}^{(t)}\mathbf{M}) \geq \mu_i^*$ holds, Lemma 9 shows that the decrement in the ℓ_1 hinge loss function value increases as $\Delta\mu_j$ increases. Hence the decrement in the ℓ_1 hinge loss function value increases as γ increases.

Third, when $\mu_i(\mathbf{R}^{(t)}\mathbf{M}) < \mu_i^*$, as γ increase, the term $\max\{\mu_i(\mathbf{R}^{(t)}\mathbf{M}) - \mu_i^*, 0\}$ remains constant, and for all j the terms $\max\{\mu_j(\mathbf{R}^{(t)}\mathbf{M}) - \mu_j^*, 0\}$ does not decrease. As γ increase, the ℓ_1 hinge loss function $L_{\mathbf{M},1}(\mathbf{R}^{(t)})$ does not decrease and thus the decrement in the ℓ_1 hinge loss function does not increase. \square

D.1 Proof of Theorem 3

The theorem follows directly from Lemma 10.

D.2 Proof of Theorems 4

It was shown in Lemma C.1 of [13] shows that with probability at least $1 - n^{-3}$ the following inequality holds:

$$\|\mathcal{P}_{\mathbf{M}}^\perp \mathbf{U}_{\widetilde{\mathbf{M}},k}\|_2 \leq C' \frac{\sigma_1(\mathbf{M})}{\sigma_k(\mathbf{M})} \sqrt{\frac{nk^2}{|\Omega|}},$$

where the projection operator $\mathcal{P}_{\mathbf{M}}^\perp = \mathbf{I} - \mathbf{M}\mathbf{M}^\dagger$. Thus we have the angular between the ranges of \mathbf{M} and $\widetilde{\mathbf{M}}$ is

$$\sin \angle(\mathbf{M}, \widetilde{\mathbf{M}}_k) \leq \|\mathcal{P}_{\mathbf{M}}^\perp \mathbf{U}_{\widetilde{\mathbf{M}},k}\|_2 \leq \frac{1}{2\rho}$$

for a large enough C defined in the theorem.

We have that for all $i, j \in [n]$, the leverage score and the cross leverage scores satisfy

$$\begin{aligned} \left\{ \begin{array}{l} |\ell_i(\mathbf{M}) - \ell_i(\widetilde{\mathbf{M}}_k)| \\ |\ell_{ij}(\mathbf{M}) - \ell_{ij}(\widetilde{\mathbf{M}}_k)| \end{array} \right\} &\leq \|\mathbf{U}_{\mathbf{M},k} \mathbf{U}_{\mathbf{M},k}^T - \mathbf{U}_{\widetilde{\mathbf{M}},k} \mathbf{U}_{\widetilde{\mathbf{M}},k}^T\|_{\max} \\ &\leq \|\mathbf{U}_{\mathbf{M},k} \mathbf{U}_{\mathbf{M},k}^T - \mathbf{U}_{\widetilde{\mathbf{M}},k} \mathbf{U}_{\widetilde{\mathbf{M}},k}^T\|_2 \\ &= \sin \angle(\mathbf{M}, \widetilde{\mathbf{M}}_k) \\ &\leq \frac{1}{2\rho}, \end{aligned}$$

from which the theorem follows.

D.3 Proof of Theorems 5

In this section we use the simplified notation defined in Table 1.

Table 1: Notation

notation	description
μ_i	the i -th leverage score of \mathbf{M}
$\hat{\mu}_i$	the approximation of μ_i with additive error at most $\pm \frac{1}{2\rho}$
$\Delta\mu_i$	equal to $\hat{\mu}_i - \mu_i$, which is the error in the estimation
μ'_i	the i -th leverage score of \mathbf{WM}
$\hat{\mu}'_i$	equal to $\frac{(1-\gamma)\hat{\mu}_i}{1-\gamma\hat{\mu}_i}$

Here and after, we omit the subscript i . Using Lemma 1, we have that the leverage scores after row weighing become

$$\begin{aligned}\mu' &= \frac{(1-\gamma)\mu}{1-\gamma\mu} = \frac{(1-\gamma)(\hat{\mu} - \Delta\mu)}{1-\gamma\hat{\mu} + \gamma\Delta\mu} \\ &= \frac{(1-\gamma)\hat{\mu}}{1-\gamma\hat{\mu}} \frac{\mu}{1-\gamma\hat{\mu} + \gamma\Delta\mu}.\end{aligned}\quad (24)$$

We define the variable

$$\hat{\mu}' = \frac{(1-\gamma)\hat{\mu}}{1-\gamma\hat{\mu}}, \quad (25)$$

which can be assigned any value in $(0, 1)$. If $\Delta\mu \ll \hat{\mu}, 1 - \hat{\mu}$, then μ' is close to $\hat{\mu}'$. Equation (25) can be equivalently written as

$$1 - \gamma\hat{\mu} = \frac{1 - \hat{\mu}}{1 - \hat{\mu}'}, \quad (26)$$

$$\gamma = \frac{1 - \hat{\mu}'/\hat{\mu}}{1 - \hat{\mu}'}. \quad (27)$$

It follows from (24) (25) (26) (27) that

$$\begin{aligned}\mu' &= \hat{\mu}' \frac{\mu}{\hat{\mu}} \frac{1 - \hat{\mu}}{1 - \hat{\mu} + \Delta\mu(1 - \hat{\mu}'/\hat{\mu})} \\ &= \hat{\mu}' \frac{\mu}{\hat{\mu}} \frac{1 - \hat{\mu}}{1 - \mu - \hat{\mu}'\Delta\mu/\hat{\mu}}.\end{aligned}$$

Taking the inverse, we have that

$$\begin{aligned}\frac{1}{\mu'} &= \frac{1}{\hat{\mu}'\mu} \left[\frac{1 - \mu}{1 - \hat{\mu}} - \frac{\Delta\mu}{1 - \hat{\mu}} \frac{\hat{\mu}'}{\hat{\mu}} \right] \\ &= \left[\frac{1}{\hat{\mu}'} - \frac{\Delta\mu}{\hat{\mu}(1 - \mu)} \right] \frac{\hat{\mu}}{\mu} \frac{1 - \mu}{1 - \hat{\mu}}.\end{aligned}\quad (28)$$

By the assumption $\hat{\mu} \in [\frac{1}{\rho}, 1 - \frac{1}{\rho}]$, we have that

$$\begin{aligned}\left| \frac{\Delta\mu}{\hat{\mu}(1 - \mu)} \right|^{-1} &= \frac{1}{|\Delta\mu|} \left| -\hat{\mu}^2 + (1 + \Delta\mu)\hat{\mu} \right| \\ &\geq \frac{1}{|\Delta\mu|} \frac{1}{\rho} \left(1 - \frac{1}{\rho} \right) + \min \left\{ \text{sgn}(\Delta\mu) \frac{1}{\rho}, \text{sgn}(\Delta\mu) \left(1 - \frac{1}{\rho} \right) \right\} \\ &\geq \left(1 - \frac{1}{\rho} \right) \left(\frac{1}{\rho|\Delta\mu|} - 1 \right).\end{aligned}$$

It follows from $|\Delta\mu| \leq \frac{1}{2\rho}$ that

$$\left| \frac{\Delta\mu}{\hat{\mu}(1 - \mu)} \right| \leq 1 + \frac{1}{\rho - 1} \approx 1.$$

Combined with (28) we have

$$\begin{aligned}\left[\frac{1}{\hat{\mu}'} - 1 - \frac{1}{\rho - 1} \right] \frac{\hat{\mu}}{\mu} \frac{1 - \mu}{1 - \hat{\mu}} \\ \leq \frac{1}{\mu'} \leq \left[\frac{1}{\hat{\mu}'} + 1 + \frac{1}{\rho - 1} \right] \frac{\hat{\mu}}{\mu} \frac{1 - \mu}{1 - \hat{\mu}}.\end{aligned}\quad (29)$$

We also have that

$$\frac{\mu}{\hat{\mu}} \frac{1 - \hat{\mu}}{1 - \mu} = 1 - \frac{\Delta\mu}{\hat{\mu}(1 + \Delta\mu - \hat{\mu})},$$

which is less than 1 if $\Delta\mu > 0$ and greater than 1 if $\Delta\mu < 0$. By the assumption $\hat{\mu} \in [\frac{1}{\rho}, 1 - \frac{1}{\rho}]$, when $\Delta\mu \geq 0$, $\frac{\mu}{\hat{\mu}} \frac{1 - \hat{\mu}}{1 - \mu}$ has a minimum of $\frac{1 - 1/\rho}{2 - 1/\rho} \approx \frac{1}{2}$ which is attained by $\hat{\mu} = 1/\rho$ and $\Delta\mu = 1/2\rho$; when $\Delta\mu \leq 0$, $\frac{\mu}{\hat{\mu}} \frac{1 - \hat{\mu}}{1 - \mu}$ has a maximum of $\frac{2 - 1/\rho}{1 - 1/\rho} \approx 2$ which is attained by $\hat{\mu} = 1 - 1/\rho$ and $\Delta\mu = -1/2\rho$. Thus we have that

$$\frac{\hat{\mu}}{\mu} \frac{1 - \mu}{1 - \hat{\mu}} \in \left[\frac{1 - 1/\rho}{2 - 1/\rho}, \frac{2 - 1/\rho}{1 - 1/\rho} \right]. \quad (30)$$

Combining (29) and (30) we have that

$$\begin{aligned}\left[\frac{1}{\hat{\mu}'} - 1 - \frac{1}{\rho - 1} \right] \frac{1 - 1/\rho}{2 - 1/\rho} \\ \leq \frac{1}{\mu'} \leq \left[\frac{1}{\hat{\mu}'} + 1 + \frac{1}{\rho - 1} \right] \frac{2 - 1/\rho}{1 - 1/\rho}.\end{aligned}$$

Setting $\frac{1}{\hat{\mu}'} = \frac{\rho - 1}{2\rho - 1} \frac{n_1}{k} - \frac{\rho}{\rho - 1} \approx \frac{n_1}{2k}$, we have that

$$\mu' \in \left[\frac{k}{n_1}, \frac{4k}{n_1} \frac{(\rho - 0.5)^2}{(\rho - 1)^2 - \frac{4k}{n_1}\rho(\rho - 0.5)} \right].$$

The scaling parameter γ is according to (27) taking $\frac{1}{\hat{\mu}'} = \frac{n_1}{2k}$ as input.

D.4 Proof of Theorems 6

We propose to set $\mu_i = \hat{\mu}_i - \frac{1}{2\rho}$ ($\leq \mu_i(\mathbf{M})$ by the additive error assumption) and $\mu'_i = \frac{1}{\rho}$, and use (12) to compute γ :

$$\gamma = \frac{1 - \mu'_i/\mu_i}{1 - \mu'_i} = \frac{\rho - \frac{1}{\hat{\mu}_i - 1/2\rho}}{\rho - 1}. \quad (31)$$

For fixed γ , the leverage score $\mu_i(\mathbf{WM})$ increases with $\mu_i(\mathbf{M})$; since $\mu_i(\mathbf{M}) \geq \mu_i = \hat{\mu}_i - \frac{1}{2\rho}$, we have $\mu_i(\mathbf{WM}) \geq \mu'_i = \frac{1}{\rho}$.

D.5 Proof of Theorem 7

By Theorems 5 and 6, the choice of γ ensures that

$$\mu_i^* \leq \mu_i(\mathbf{R}^{(t)}\mathbf{M}) < \mu_i(\mathbf{R}^{(t-1)}\mathbf{M}).$$

Thus the decrement in the objective function follows directly from Lemma 9.

When \mathcal{J} is non-empty, for all $j \in \mathcal{J}$, $\Delta\mu_j > 0$ because $\mu_{ij}(\mathbf{R}^{(t-1)}\mathbf{M}) > 0$ and $\mu_i(\mathbf{R}^{(t-1)}\mathbf{M}) - \mu_i(\mathbf{R}^{(t)}\mathbf{M}) > 0$. That \mathcal{J} is non-empty also ensures that $\mu_j(\mathbf{R}^{(t-1)}\mathbf{M}) < \mu_i^*$. Hence the decrement in the objective function (17) is strictly greater than zero.

D.6 Proof Corollary 8

The theorem follows directly from Lemma 10.

D.7 Leverage Scores

THEOREM 11. *Let \mathbf{A} be an $n_1 \times n_2$ matrix of rank r ($< n_1, n_2$), $\mathbf{B} \in \mathbb{R}^{n_1 \times r}$ be arbitrary bases of \mathbf{A} , and \mathbf{R} be any $n_1 \times n_1$ nonsingular matrix. Then the leverage scores of \mathbf{RA} and \mathbf{RB} are the same.*

PROOF. Let $\mathbf{U}_\mathbf{A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}^T$ and $\mathbf{U}_\mathbf{B}\Sigma_\mathbf{B}\mathbf{V}_\mathbf{B}^T$ be the condensed SVD of \mathbf{A} and \mathbf{B} , respectively. Since \mathbf{B} is the bases of \mathbf{A} , there exists an $r \times n_2$ matrix \mathbf{C} such that $\mathbf{A} = \mathbf{BC}$. We let $\mathbf{D} = \Sigma_\mathbf{B}\mathbf{V}_\mathbf{B}^T\mathbf{C} \in \mathbb{R}^{r \times n_2}$ and obtain

$$\begin{aligned}\mathbf{A} &= \mathbf{BC} = \mathbf{U}_\mathbf{B}\Sigma_\mathbf{B}\mathbf{V}_\mathbf{B}^T\mathbf{C} \\ &= \mathbf{U}_\mathbf{B}\mathbf{D} = (\mathbf{U}_\mathbf{B}\mathbf{U}_\mathbf{D})\Sigma_\mathbf{D}\mathbf{V}_\mathbf{D}^T.\end{aligned}$$

It is not hard to see that $\mathbf{U}_\mathbf{B}\mathbf{U}_\mathbf{D} = \mathbf{U}_\mathbf{A} \in \mathbb{R}^{n_1 \times r}$. Let $\mathbf{u}_\mathbf{A}^{(i)}$ and $\mathbf{u}_\mathbf{B}^{(i)}$ be the i -th row of \mathbf{A} and \mathbf{B} , respectively. Since $\mathbf{U}_\mathbf{D}$ is an $r \times r$ orthogonal matrix, we have that

$$\|\mathbf{u}_\mathbf{A}^{(i)}\|_2^2 = \|\mathbf{u}_\mathbf{B}^{(i)}\mathbf{U}_\mathbf{D}\|_2^2 = \|\mathbf{u}_\mathbf{B}^{(i)}\|_2^2$$

for all $i \in [n_1]$. The theorem follows by the definition of leverage scores. \square